Multiple testing:
False discovery rate and the q-value

Stefanie Scheid

Max-Planck-Institute for Molecular Genetics

Computational Diagnostics

June 17, 2002

Articles

- Storey, J.D. (2001a): **The positive False Discovery Rate: A Bayesian Interpretation and the q-value**, submitted
- Storey, J.D. (2001b): **A Direct Approach to False Discovery Rates**, submitted
- Storey, J.D., Tibshirani, R. (2001): **Estimating False Discovery Rates Under Dependence, with Applications to DNA Microarrays**, submitted
- `http://www-stat.stanford.edu/~jstorey/`

The multiple testing problem

- random variables $X_1, \ldots, X_k$ from same family of distribution $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$
- set of m hypotheses for θ
- conclude from **one** sample

Why not testing separately?

m independent test statistics with level α

Why not testing separately?

m independent test statistics with level α

$\Rightarrow$ Prob$(\textit{at least 1 is falsely rejected}) = 1 - (1-\alpha)^m = \alpha_m$

Why not testing separately?

m independent test statistics with level α

$\Rightarrow$ Prob$(\textit{at least 1 is falsely rejected}) = 1 - (1-\alpha)^m = \alpha_m$

$\Rightarrow$ control α_m with multiple test procedure

Two kinds of multiple testing procedures: One-step

- test each null hypothesis "independently" from outcome of others
- e.g. Bonferroni: test each hypothesis to level $\frac{\alpha}{m}$

$\Rightarrow$ multiple level $= \alpha$

Two kinds of multiple testing procedures: Multi-step

- test in sorted order dependent on outcome
- e.g. Bonferroni-Holm: sort according to p-values and test with increasing α: $\frac{\alpha}{m}, \frac{\alpha}{m-1}, \ldots, \alpha$

$\Rightarrow$ multiple level $= \alpha$

Microarray data

- k samples (e.g. in 2 groups) and m genes

Microarray data

- k samples (e.g. in 2 groups) and m genes
- observe gene expression as intensity values:

	sample 1	sample 2	...	sample k
gene 1	404	1873	...	151
gene 2	-2015	-716	...	1227
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
gene m	126	42	...	85

Number of outcomes

	Not rejected	Rejected	Σ
Null true	U	V	m_0
Alternative true	T	S	m_1
Σ	W	R	m

Problems

- multiple test controls Prob$(V \geq 1) \leq \alpha$

Problems

- multiple test controls Prob$(V \geq 1) \leq \alpha$
- m is huge $\Rightarrow$ falsely rejected are likely to occur

Problems

- multiple test controls Prob$(V \geq 1) \leq \alpha$
- m is huge $\Rightarrow$ falsely rejected are likely to occur
- better control $\dfrac{\sharp\{falsely\ rejected\}}{\sharp\{rejected\ in\ total\}}$

Problems

- multiple test controls Prob$(V \geq 1) \leq \alpha$
- m is huge $\Rightarrow$ falsely rejected are likely to occur
- better control $\dfrac{\sharp\{falsely\ rejected\}}{\sharp\{rejected\ in\ total\}}$
- intuitive definition of false discovery rate:

$$FDR = \mathsf{E}\left[\frac{V}{R}\right]$$

Difference between multiple testing and FDR

- **Multiple testing:**

  Fixed error rate $\Rightarrow$ estimated rejection area

- **FDR:**

  Fixed rejection area $\Rightarrow$ estimated error rate

What if $R = 0$?

- Benjamini and Hochberg: $FDR = \mathsf{E}\left[\frac{V}{R} \mid R > 0\right] \cdot \mathsf{Prob}(R > 0)$

  "the rate that false discoveries occur"

What if $R = 0$?

- Benjamini and Hochberg: $FDR = \mathsf{E}\left[\frac{V}{R} \mid R > 0\right] \cdot \mathsf{Prob}(R > 0)$

  "the rate that false discoveries occur"

- Storey: $pFDR = \mathsf{E}\left[\frac{V}{R} \mid R > 0\right]$

  "the rate that discoveries are false"

FDR in Bayesian terms

Theorem: m identical hypothesis tests are performed with independent statistics $T_1, \ldots, T_m$ and rejection area C. A null hypothesis is true with a-priori probability $\pi_0 = \text{Prob}(H = 0)$. Then

$$pFDR(C) = \frac{\pi_0 \cdot \text{Prob}(T \in C \,|\, H = 0)}{\text{Prob}(T \in C)} = \text{Prob}(H = 0 \,|\, T \in C).$$

Algorithms for calculating $\widehat{FDR}$ and $\widehat{pFDR}$ in Storey (2001b).

Simulation study

1. independence between genes
2. loose ("clumpy") dependence
3. general dependence

Simulation study

1. independence between genes
2. loose ("clumpy") dependence
3. general dependence

1. + 2. $\widehat{pFDR}$ is very accurate

3. $\widehat{pFDR}$ is biased upward (overestimation of π_0)

Reasons for "clumpy dependence" in microarrays

- genes work in pathways
  $\Rightarrow$ small groups of genes interact to produce overall process

Reasons for "clumpy dependence" in microarrays

- genes work in pathways

  $\Rightarrow$ small groups of genes interact to produce overall process

- cross-hybridization

  $\Rightarrow$ genes with molecular similarity have evolutionary and/or functional relationship

p-value vs. q-value

- **Definition:** For a nested set of rejection areas $\{C\}$ define the $p\text{-}value$ of an observed statistic $T = t$ to be:

$$p\text{-}value(t) = \min_{\{C:\, t \in C\}} \text{Prob}(T \in C \,|\, H = 0).$$

p-value vs. q-value

- **Definition:** For a nested set of rejection areas $\{C\}$ define the $p\text{-}value$ of an observed statistic $T = t$ to be:

$$p\text{-}value(t) = \min_{\{C:\, t \in C\}} \text{Prob}(T \in C \,|\, H = 0).$$

- **Definition:** For an observed statistic $T = t$ define the $q\text{-}value$ of t to be:

$$q\text{-}value(t) = \min_{\{C:\, t \in C\}} pFDR(C) = \min_{\{C: t \in C\}} \text{Prob}(H = 0 \,|\, T \in C).$$

"posterior Bayesian p-value"

Conclusion

- FDRs are more appropriate in large sets of hypotheses than multiple testing procedures

Conclusion

- FDRs are more appropriate in large sets of hypotheses than multiple testing procedures
- one has to be sure about rejection area

Conclusion

- FDRs are more appropriate in large sets of hypotheses than multiple testing procedures
- one has to be sure about rejection area
- positive FDR ("rate that discoveries are false") is more appropriate than FDR ("rate that false discoveries occur")

Conclusion

- FDRs are more appropriate in large sets of hypotheses than multiple testing procedures
- one has to be sure about rejection area
- positive FDR ("rate that discoveries are false") is more appropriate than FDR ("rate that false discoveries occur")
- q-value can be reported with every statistic ("minimum pFDR over which that statistic can be rejected")